Kataloganreicherung und Zeitschriftenerschließung mit MyBib eDoc® und C-3 am Ibero-Amerikanischen Institut, Preußischer Kulturbesitz

Neue Verfahren zur Optimierung der bibliografischen Nachweissituation in einer großen Spezialbibliothek

Christoph Müller, Nicolai Sternitzke, Rüdiger Stratmann, Thomas Parschik

Die Bibliothek des Ibero-Amerikanischen Instituts der Stiftung Preußischer Kulturbesitz in Berlin (IAI) ist mit über 830.000 Monografien, 33.000 Zeitschriften und Zeitungen, von denen ca. 5.000 als laufende Abos geführt werden, und zahlreichen weiteren großen Sondersammlungen (Landkarten, Tonträger, Videos, DVDs, Nachlässe etc.) die größte Spezialbibliothek zu Lateinamerika, Spanien, Portugal und der Karibik in Europa und nach der Library of Congress in Washington und der Nettie-Lee-Benson-Collection der University of Texas in Austin die drittgrößte Spezialbibliothek dieser Art in der Welt. Über Kauf, Tausch und Schenkung erweitert sich der Bestand jedes Jahr um ca. 30.000 Monografien.

Um die Nachweissituation von Sammelband- und Zeitschriftenaufsätzen und gleichzeitig das bibliografische Informationsangebot im Sinne einer Spezialbibliothek zu verbessern, hat das IAI in den letzten zwei Jahren zwei neue Geschäftsgänge eingeführt, mit denen die seit der Gründung des IAI vor fast 80 Jahren gängige Praxis, ausgewählte Aufsätze aus Zeitschriften und Sammelbänden zu erschließen, automatisiert wurde.

Als erstes trat die Bibliothek des IAI 2007 dem System der Online-Contents-Sondersammelgebietsausschnitte (OLC-SSG) des GBV bei und betreut seitdem den OLC-SSG Ibero-Amerika. Es werden dazu im IAI sowohl die aktuellen als auch die bis zum Jahr 2000 zurückreichenden Inhaltsverzeichnisse von 780 laufenden Zeitschriftentiteln auf Artikelebene erschlossen.

Zur Bewältigung dieser Menge an Artikeldaten kommt im IAI die Software C-3 der ImageWare Components GmbH zum Einsatz, die es ermöglicht, die formale Erschließung in einem größtenteils automatisierten Geschäftsgang durchzuführen. In den einzelnen Modulen von C-3 erfolgt die automatische Erkennung der Titel-, Autoren- und Seitenzahlinformationen in den gescannten Inhaltsverzeichnissen sowie deren automatische Indexierung und Konversion in Katalogisate.

An den Indexierungsarbeitsplätzen kommen normale Standard-PCs zum Einsatz, über die auf die Auftragsverwaltung des zentralen C-3 Periodikaservers per Browser zugegriffen wird.

Die C-3 Software besteht aus den Programmmodulen C-3 Template und C-3 Index. In C-3 Template werden einmalig für jeden Zeitschriftentitel die Struktur der Inhaltsverzeichnisse (Interpretationstyp Regel, Tabelle oder Freiform), die Abfolge von Aufsatztitel, Autoren und Seitenzahlen sowie die Schriftattribute (fett, kursiv etc.) der bibliografischen Daten festgelegt.

Auf Basis dieser Templatedefinition werden die Scans der jeweiligen Inhaltsverzeichnisse im C-3 Index Modul mit der OCR Software Abbyy Fine Reader so erkannt, dass die bibliografischen Informationen der einzelnen Aufsätze bereits kategorisiert und aufsatzweise separiert ausgegeben werden. Die dabei erzeugten Daten können, sofern erforderlich, noch während der Bearbeitung mit C-3 mit einer Reihe von integrierten Nachbearbeitungstools korrigiert und im Anschluss im XML-Format an den C-3 Periodikaserver exportiert werden. Automatische Konversionsroutinen erzeugen aus den xml-Dateien serverseitig Artikeldaten im Pica3-Format, die nach einer abschließenden Qualitätskontrolle in die Online Contents Datenbanken des GBV eingespielt werden.

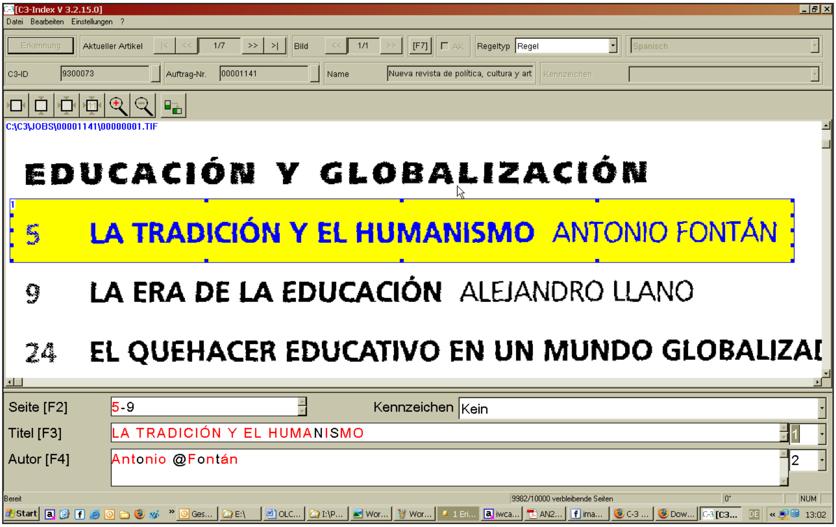


Abb. 1: Bearbeitungsoberfläche in C-3 Index

Um die Aufsätze der jeweiligen Zeitschrift eindeutig zuordnen zu können, werden die Titelaufnahmen der Zeitschriften, die Templates und später auch die jeweiligen Aufsatztiteldaten über eine eindeutige Ident-Nummer, die C-3 ID, verknüpft. Dazu bietet sich neben der Pica-Produktionsnummer auch die so genannte

Swets-Nummer an. Das OLC-SSG-System des GBV beruht auf der Swets-Datenbank Online Contents, in der alle von Swets gelieferten Aufsatzdaten verzeichnet und recherchierbar sind.

Die Steuerung des Gesamtworkflows erfolgt über den von der Verbundzentrale des GBV in Göttingen gehosteten MyBib eDoc®-basierten C-3 Periodikaserver, über den die Mitarbeiter jederzeit den Status des Bearbeitungsprozesses im Geschäftsgang nachverfolgen können. Der C-3 Periodikaserver bietet offene Schnittstellen zur Anbindung aller für den Geschäftsgang erforderlichen Workflowkomponenten und ermöglicht über seine webbasierte mehrplatzfähige Auftragsverwaltung eine effiziente Steuerung der Scan- und Indexierungsprozesse.

Die Erzeugung eines Scan- und Indexierungsauftrages auf dem C-3 Periodikaserver erfolgt automatisch bei der Akzession eines eingehenden Zeitschriftenheftes im Erwerbungsmodul von LBS4 (ACQ). Dabei werden die bibliografischen Daten des zu erschließenden Zeitschriftenheftes mittels eines von der VZG entwickelten Zusatzprogramms von einem in LBS4 erzeugten elektronischen Belegzettel ausgelesen und in eine ILL-subito konforme Bestellmail umgewandelt. Durch Versendung dieser Bestellmail an den C-3 Periodikaserver wird ein Scan- und Indexierungsauftrag zum akzessionierten Heft erzeugt, der die bibliografischen Metadaten des Auftrages mit dem gescannten Inhaltsverzeichnis und den Indexierungsergebnissen verknüpft. Damit werden alle zum Auftrag gehörenden Daten übersichtlich und nachvollziehbar verwaltet und können bei Bedarf von den Mitarbeitern editiert werden.

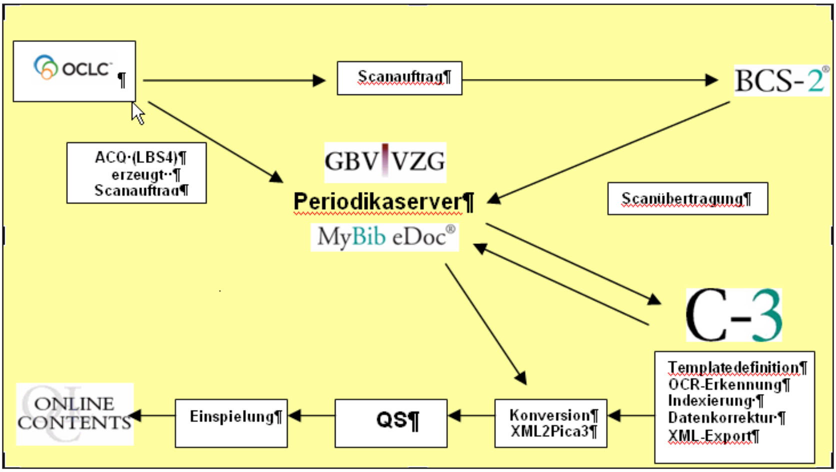


Abb.2: Workflowskizze C-3 Periodikaserver

Monatlich entstehen seitdem mehr als 3.000 Aufsatztiteldatensätze aus aktuellen und retrospektiv erschlossenen Zeitschrifteninhaltsverzeichnissen, die in den GVK+ und ab März 2010 auch in den OPAC des IAI eingespielt werden. Die Erkennungsqualität des Unicode-fähigen C-3 Verfahrens ist selbst bei komplexen, mehrsprachigen Inhaltsverzeichnissen mit zahlreichen diakritischen Zeichen so hoch, dass Qualitätskontrollen und Datenkorrekturen nur in kleinem Rahmen vorgenommen werden müssen. Das bedeutet, dass neben der Auftragserzeugung auch die Indexierung und die Qualitätskontrolle der produzierten Artikeldaten im laufenden Betrieb durchgeführt werden können.

Für das Scannen der Inhaltsverzeichnisse kommen ca. 0,3 Vollzeitäquivalente und bei der Indexierung ein VZÄ zum Einsatz. Die Abschlusskorrektur zur Sicherung der Datenqualität liegt am IAI in der Verantwortung eines Diplom-Bibliothekars, alle anderen Arbeitsschritte werden von Assistenten bzw. Fachangestellten für Medien- und Informationsdienste ausgeführt.

Insgesamt hat sich das System als sehr effizient erwiesen. Der Produktionsbetrieb verläuft störungsfrei und die Bedienung der Software ist komfortabel. Durch den hohen Automatisierungsgrad wird die Katalogisierung von Zeitschriftenaufsätzen für den SSG Online Contents-Dienst sehr erleichtert und beschleunigt, so dass in den letzten zwei Jahren insgesamt 51.000 Artikeldatensätze produziert werden konnten, die über den GVK+ und in wenigen Wochen auch über den OPAC der Bibliothek des IAI zugänglich sind.

Im Bereich der nicht periodisch erscheinenden Publikationen werden seit Mai 2009 die Inhaltsverzeichnisse und Cover von Monografien, die neu vom IAI erworben werden, gescannt und als durchsuchbare pdf-Dateien bzw. Image-Dateien an die jeweiligen Katalogeinträge im CBS des GBV angehängt und damit im OPAC des IAI sowie im Verbundkatalog angezeigt.

In diesem Geschäftsgang entscheiden die FachreferentInnen des IAI während der Inhaltserschließung der Bücher, ob die jeweiligen Inhaltsverzeichnisse geeignet sind, den Katalogeintrag mit weiteren Informationen anzureichern. Ausgewählt werden dabei zum einen die Inhaltsverzeichnisse von Sammelbänden, da die darin verzeichneten Aufsatzinformationen durch die Volltextindexierung der OCR-Daten recherchierbar gemacht werden und damit den Band wesentlich tiefer erschließen. Zum anderen werden Monografien in die weitere Bearbeitung gegeben, deren Inhaltsverzeichnisse Kapitelüberschriften enthalten, die in Ergänzung zur intellektuellen Schlagworterschließung weitere Detailinformationen zum Inhalt der Monografie liefern. Die zum Scannen ausgewählten Cover sollen weitere zusätzliche Informationen zur Publikation liefern und den OPAC für den Benutzer attraktiver machen. Das Benutzerfeedback auf den Coverdienst ist durchweg positiv.

Die Inhaltsverzeichnisse der ausgewählten Bände werden bei diesem Verfahren an einem speziell ausgestatteten Arbeitsplatz gescannt und durchlaufen anschließend eine OCR-Texterkennung, sodass die Bände unmittelbar nach der Bearbeitung wieder in den Standardgeschäftsgang zurückgeführt werden können.

Für den Scanprozess wird ein WideTekA2-Flachbettscanner verwendet, der mit dem Scan-Client BCS-2® betrieben wird. Der Scan-Client mit integrierter OCR-Funktion ist an das Workflowsteuerungssystem MyBib eDoc® angebunden und gewährleistet über diverse Bildnachbearbeitungsfunktionen die Scanqualität der erzeugten Digitalisate.

Die Inhaltsverzeichnisse werden als bitonale Images im TiFF-Format mit einer Auflösung von 300dpi gescannt, um optimale Voraussetzungen für eine gute OCR-Erkennungsrate zu bieten. Die Buchcover werden dagegen in Farbe gescannt und im JPG-Format gespeichert. Um die unterschiedlichen Formate und Auflösungen zu erzeugen, sind im verwendeten Scan-Client BCS-2® zwei Auftragstypen mit entsprechenden Scanprofilen konfiguriert, zwischen denen flexibel umgeschaltet werden kann.

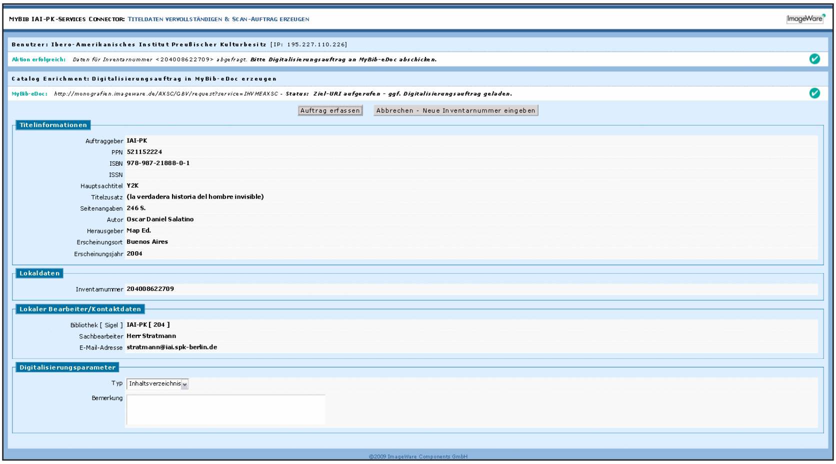


Abb.3: MyBib eDoc® Bestellformular

Um für diesen Geschäftsgang einen Digitalisierungsauftrag zu einem ausgewählten Band zu erzeugen, wird über ein Webformular mit Hilfe eines Barcodelesers die Inventarisierungsnummer des Bandes eingelesen. Nach Absenden des Formulars wird eine Abfrage an den IAI Web-OPAC angestoßen und damit die Auftragsdaten in einer MyBib eDoc® Bestellmaske vervollständigt und angezeigt. Mit Auslösen der Bestellung werden die zugehörigen Metadaten des Titels sowie die

eindeutige Kennung im Verbundsystems (PPN) an MyBib eDoc® übertragen und in Folge der Scanauftrag generiert. Durch eine zwischengeschaltete Dublettenkontrolle wird anhand der PPN ferner überprüft, ob im Verbundkatalog bereits ein Inhaltsverzeichnis zum Titel vorliegt. Die Auftragserzeugung wird in diesem Falle automatisch unterbunden.

Nach der Auftragserzeugung wird am Scanarbeitsplatz ein Auftragszettel mit allen relevanten bibliografischen Daten ausgedruckt, der sowohl die eindeutige MyBib eDoc® Auftragsnummer als auch die Pica-Produktionsnummer (PPN) des Titels im Verbundsystem als Barcode beinhaltet. Durch Einscannen des Auftragszettels wird mittels einer automatischen Barcodeerkennung der Scanauftrag von BCS-2® auf dem MyBib eDoc®-Server aufgerufen und der Bearbeitungsstatus serverseitig fortgeschrieben. Im Anschluss wählt der Auftragsoperator über eine für die Bedürfnisse des IAI konfigurierte Spezialtastatur die Sprache der Vorlage aus und kann damit das hinterlegte Wörterbuch für die OCR-Erkennung auswählen. Nach Scanning des Inhaltsverzeichnisses löst der Scanoperator über die Sondertastatur die in BCS-2 integrierte OCR-Funktion (Abbyy Fine Reader Engine) aus und kann in der BCS-2® Oberfläche die OCR-Erkennungsergebnisse mit dem Scan des Inhaltsverzeichnisses vergleichen. Eine Datenkorrektur ist aufgrund der nahezu fehlerfreien Erkennungsrate nicht erforderlich. Anschließend kann das Digitalisat gemeinsam mit der OCR-Datei an MyBib eDoc® hochgeladen werden. Die Anreicherung der Katalogisate im GVK und im OPAC des IAI liegt in der Verantwortung der Verbundzentrale des GBV. Dafür werden die gewonnenen Digitalisate und die zugehörigen OCR-Dateien vom MyBib eDoc®-Server zunächst in ein Repository (MyBib TOX) geladen. In einem zweiten Schritt können die verlinkten Dateien über einen verfügbaren Webservice von der Verbundzentrale abgeholt und in regelmäßig laufenden Prozessen in den GVK und in den IAI-OPAC eingespielt werden. Dabei wird im Verbundkatalog und im IAI Web-OPAC ein Link zum Aufruf des Inhaltsverzeichnisses integriert, der zum durchsuchbaren PDF-Dokument führt.

Die Digitalisierung der Cover wird in einem etwas abweichenden Verfahren vollzogen. Dafür wird im Scan-Client BCS-2 ein Scanauftrag mit der PPN des Verbundkatalogisates angelegt, im JPEG-Format gescannt und das Digitalisat im Anschluss per FTP in ein Repository der Verbundzentrale geladen. Durch die eindeutige PPN, die aus der Dateinamenskonvention hervorgeht, können die Coverscans bei der Kataloganreicherung eindeutig mit dem Titeldatensatz im Verbundkatalog in der PICA-Kategorie 4099 verknüpft und angereichert werden. Das Cover wird in der Titelanzeige als Thumbnail visualisiert und kann durch Anklicken in mehreren Stufen vergrößert werden.

Eine gezielte Suche in den Inhaltsverzeichnissen im OPAC ist möglich. Mit Hilfe des Suchschlüssels „txt" kann direkt eine Katalogrecherche über die im OPAC angereicherten Titelaufnahmen durchgeführt werden. Als Suchergebnis werden alle

Bände, deren Inhaltsverzeichnisse den Recherchebegriff enthalten, aufgelistet. Um überprüfen zu können, ob der Titel auch in den gesuchten inhaltlichen Zusammenhang passt, werden sowohl in der Kurzliste als auch in der vollständigen Anzeige der Titeldaten auszugsweise die Phrasen mit dem Suchbegriff aus den Inhaltsverzeichnissen angezeigt.

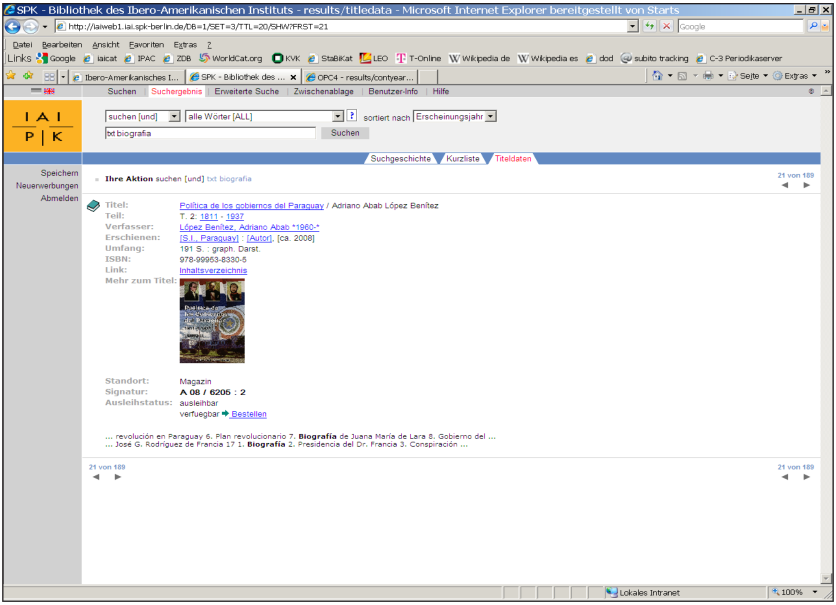


Abb. 4: IAI Web-OPAC Titelanzeige mit Link zum Inhaltsverzeichnis und Cover-Anzeige

Nach der Installation des Systems, einer ausführlichen Schulung und einer Testphase im 2. Quartal 2009 läuft seit Juni 2009 der Produktionsbetrieb. In den ersten acht Monaten wurde der Katalog dabei um ca. 4.200 Inhaltsverzeichnisse und ca. 5.200 Cover angereichert. Die Produktion hat sich inzwischen aber noch deutlich erhöht.

Es ist bisher aus technischen Gründen zwar noch nicht möglich, Cover und Inhaltsverzeichnis eines Bandes in einem Schritt zu bearbeiten, so dass die Scanprozesse für beide Anreicherungsobjekte in separaten Bearbeitungsschritten erfolgen müssen. Das System ist aber so schnell und komfortabel konzipiert, dass für die Bearbeitung der oben genannten Menge an Inhaltsverzeichnissen und Covern nur ca. 0,3 Vollzeitäquivalente eines Bibliotheksassistenten benötigt werden. Das System ist darüberhinaus so bedienerfreundlich, dass Mitarbeiter mit geringen

Vorkenntnissen und wenig Scanpraxis (z.B. Praktikanten) sehr schnell in die intuitive Handhabung des Systems eingearbeitet werden können. Durch diese getrennten Arbeitsschritte hat sich in der Praxis durchgesetzt, dass stapelweise erst die Scans der Cover der Monografien gemacht werden und dann die Inhaltsverzeichnisse bearbeitet werden. Insgesamt hat sich gezeigt, dass unter Verwendung der beschriebenen Systeme mit vergleichsweise wenig Aufwand sehr effektive Geschäftsgänge zur Kataloganreicherung und Zeitschrifteninhaltserschließung eingeführt werden konnten. Diese ermöglichen es, in kurzer Zeit und substantiellen Volumen die Katalognachweise von Zeitschriften und Monografien aus den Beständen der Bibliothek des IAI nachhaltig zu ergänzen. Damit wird neben der Verbesserung der Katalogqualität mit erweiterten Recherchemöglichkeiten auch insbesondere der Service für die Benutzerinnen und Benutzer des Ibero-Amerikanischen Institutes verbessert. Selbstverständlich sind alle durch das IAI im Rahmen des Catalogue Enrichment erfassten Daten durch die Partnerbibliotheken des GBV nachnutzbar.